

A brief overview of the sock matching problem

Bojana Pantić^a, Olga Bodroža-Pantić^a

^aDept. of Math. & Info., Faculty of Science, University of Novi Sad, Novi Sad, Serbia

Abstract.

This short note deals with the so-called *Sock Matching Problem* which appeared in [S. Gilliland, C. Johnson, S. Rush & D. Wood, The sock matching problem, *Involve*, 7 (5) (2014), 691–697.]. Let us denote by $B_{n,k}$ the number of all the sequences a_1, \dots, a_{2n} of nonnegative integers with $a_1 = 1$, $a_{2n} = 0$ and $|a_i - a_{i+1}| = 1$ containing at least one number k ($1 \leq k \leq n$). The value a_i can be interpreted as the number of unmatched socks being present after drawing the first i socks randomly out of the pile which initially contained n pairs of socks. Here, establishing a link between this problem and with both some old and some new results, related to the number of restricted Dyck paths, we prove that the probability for k unmatched socks to appear (in the very process of drawing one sock at a time) approaches 1 as the number of socks becomes large enough. Furthermore, we obtain a few valid forms of the sock matching theorem, thereby at the same time correcting the omissions made in the above mentioned paper.

1. Introduction

In simple terms, what is understood under The Sock Matching Problem [2] is the following procedure. Out of the laundry pile that contains exactly n different pairs of socks socks are being drawn randomly, one at a time (so that in the end all the $2n$ socks would get matched). In each move one tries to find the adequate pair among the drawn socks, in case it had already been obtained in the process. Furthermore, each of the two options: drawing a match to some sock or drawing a sock that has no match as of yet matches a single move, either one unit up or one unit to the right, on a Dyck path (a path in an $n \times n$ grid starting from the lower left corner $(0, 0)$ and ending in the upper right corner (n, n) using merely moves up and to the right without ever crossing the diagonal, see Fig. 1a) - this particular model of the Dick paths was used in [2]).

It is a well known fact indeed that the number of all the Dick paths of order n is equivalent to the n^{th} Catalan number $C_n = \frac{1}{n+1} \binom{2n}{n}$ (see [7]).

Let us now formulate our Sock Matching Problem in somewhat mathematically stricter terms. In fact, let us focus upon the total number of ways, labeled by $B_{n,k}$, to get at least k unmatched socks at least once during the matching process. Considering the aforementioned interpretation of our problem using Dick paths, it is our task to determine which ones out of these C_n possibilities are those that present the paths which hit or pass above the line $y = x + k$.

Admittedly there is a wide range of interpretations of the Catalan numbers C_n . However, it is that which allows various authors to opt themselves for the most suitable one. Here, we make use of the following terminology from [7, 8]:

2010 *Mathematics Subject Classification*. Primary 05A15, 05A16; Secondary 03B48, 00A69

Keywords. Sock matching; Dyck path; generating functions

Research supported by the Ministry of Education and Science of the Republic of Serbia (Grants OI 174018 and III 46005)

Email addresses: dmi.bojana.pantic@student.pmf.uns.ac.rs (Bojana Pantić), olga.bodroza-pantic@dmi.uns.ac.rs (Olga Bodroža-Pantić)

- Lattice paths, used in [4], consider the up- and down-steps. The former $(1, 1)$ -steps represent the case when "a sock with no match has been drawn", whereas the latter $(1, -1)$ -steps represent the case when "a match has been made". Here, a lattice path goes from $(0, 0)$ to $(2n, 0)$ on the Cartesian plane without ever moving across the x-axis (though it is allowed to hit it), as shown in Fig. 1b);
- The number of planted plane trees (i.e. rooted trees which have been embedded in the plane so that the relative order of subtrees at each branch is part of its structure; ordered trees) with $n + 1$ nodes, see Fig. 2;
- Discrete random walks in a straight line with an absorbing barrier at 0 (represented by the sequences $1 = c_1, c_2, c_3, \dots, c_{2n+2} = 0$; where $c_i \geq 1$ for $i < 2n + 2$ and $|c_i - c_{i+1}| = 1$);
- The number of all the sequences a_1, \dots, a_{2n} of nonnegative integers with $a_1 = 1, a_{2n} = 0$ and $|a_i - a_{i+1}| = 1$ (Problem 6.19 (u^5) in [8]). (Note that the value a_i can be interpreted as the exact number of unmatched socks which are present after drawing the i^{th} sock. Let us note that further on we shall refer to this interpretation as the one with the nonnegative sequences).

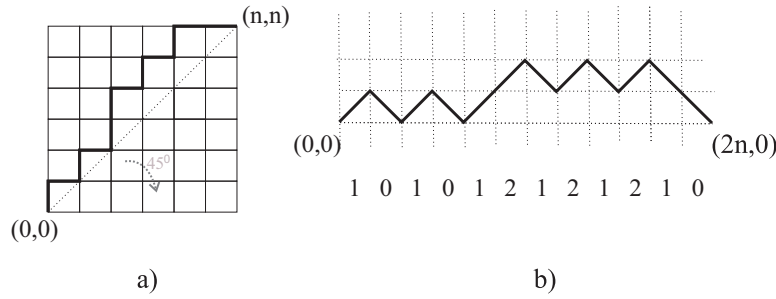


Figure 1: a) a Dyck path b) a lattice path with a corresponding nonnegative sequence

A bijection is easily established between the sets of the aforementioned combinatorial objects. For instance, the Dyck path shown in Fig. 1a) corresponds to both the path in Fig. 1b) (which can be obtained by rotating that very figure for -45° and then expanding it with the expansion coefficient of $\sqrt{2}$) as well as to the tree in Fig. 2. To be more precise, by wandering around that tree the vertical component of successive positions describes a path from 1 (the root of the tree) to 0. Consequently, in this particular example the corresponding discrete random walk would be 1, 2, 1, 2, 1, 2, 3, 2, 3, 2, 3, 2, 1, 0; whereas the nonnegative sequence, mentioned in the last interpretation, would be 1, 0, 1, 0, 1, 2, 1, 2, 1, 2, 1, 0.

The height of a Dyck path is the greatest distance from the diagonal to the path, the height of a lattice path the greatest distance from the x-axis to the path, whilst the height of a planted ordered tree is the number of nodes on a maximal simple path starting at a root. Clearly, the height of a Dyck path is $\max_i a_i$ in the nonnegative sequences interpretation, whereas the height of a corresponding planted plane tree is $\max_i c_i$ in the discrete random walks interpretation. Now, it is worthwhile realising that the latter value is for one greater than the former. To illustrate this point, do have another look at Figures 1 and 2. There, the height of the presented Dyck path (Fig. 1) is 2 (at most 2 unmatched socks appear), as opposed to the height of the planted plane tree (Fig. 2) which is 3!

Bearing all this in mind, the outlined problem from the heading, i.e. the number $B_{n,k}$ may represent as follows:

- the number of Dyck paths of height at least k (the ones that hit or cross the line $y = x + k$),
- the number of lattice paths of height at least k (the ones that hit or cross the line $y = k$),
- the number of planted plane trees of height at least $k + 1$,

- the number of discrete random walks with $\max_i c_i \geq k + 1$,
- the number of all the nonnegative sequences containing the letter k .

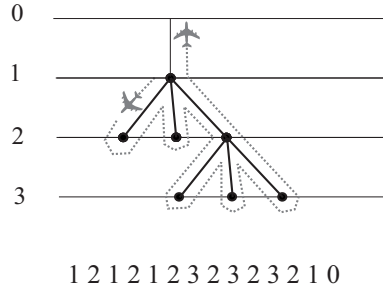


Figure 2: a tree with its random walk

Contemporary research related to the Dyck paths refer mainly to the number of restricted lattice paths, where the crossing of the x-axis is allowed. Ilić and Ilić in [4] gave the upper and lower bounds for this number in the form of binomial coefficients. Forging a link between this problem and an older paper [5] from 1985. H. Prodinger in [6] provides an explicit formula for those, as given in this theorem:

Theorem 1.1. *The number of random walks from $(0, 0)$ to $(2n, 0)$ with up-steps and down-steps of one unit each, under the condition that the path is placed between the lines $y = -h$ and $y = t$ is equal to*

$$\sum_{j \geq 0} \left[\binom{2n}{n - j(h + t + 2)} - \binom{2n}{n - j(h + t + 2) - h - 1} - \binom{2n}{n - j(h + t + 2) - t - 1} + \binom{2n}{n - (j + 1)(h + t + 2)} \right]. \quad (1)$$

In a special case, for $h = 0$, we obtain the number of all the sequences a_1, \dots, a_{2n} of nonnegative integers with $a_1 = 1, a_{2n} = 0$ and $|a_i - a_{i+1}| = 1$ and $a_i \leq t$ ($n \geq 1, t \geq 0$). Let us label it by $A_{n,t}$. Obviously, $A_{n,0} = 0$ and $A_{n,1} = 1$ for $n \geq 1$, $A_{n,t} = C_n$ for $t \geq n$. The value $A_{n,t}$ was already essentially obtained in [1] in the distant 1972. in the form of trigonometric functions. The authors of that paper used the rooted tree (planted plane tree) interpretation. As for convenience, we reformulate their results in the following theorem.

Theorem 1.2.

$$A_{n,t} = \frac{1}{t+2} \sum_{1 \leq j \leq \frac{t+1}{2}} 4^{n+1} \sin^2 \left(\frac{j\pi}{t+2} \right) \cos^{2n} \left(\frac{j\pi}{t+2} \right), \quad n \geq 1 \quad (2)$$

It is quite an interesting fact that this formula has been rediscovered many a times, which the authors of the paper [1] clearly point out, and that above all Lagrange derived a formula in 1775. which essentially includes this as a special case.

The authors of the paper [2] derive a recurrence formula for the numbers $B_{n,k}$ to which they refer as the so-called *Sock Matching Theorem*. However, we noticed that an omission was made in the proof of it having written “the first point” instead of “the last point” whilst defining the point (i, i) . Additionally, the index of summation should go from $i = 0$ to $i = n - 1$, and NOT from $i = 1$ to $i = n$. Further, having found the results from the included Table 1. to be correct, it is pretty clear that the paper must have contained a few typos. For the purpose of providing a valid formula for $B_{n,k}$, with the accompanying proofs, we shall present two equivalent alternatives in Section 2, using more than just the mentioned author’s idea. In section 3 we give an explicit expression for $B_{n,k}$.

In [2] a proposition was made that the probability $P_{n,k}$ for the Dyck path to reach the line $y = x + k$ approaches 1 as the number n becomes large enough, i.e. $\lim_{n \rightarrow \infty} \frac{B_{n,k}}{C_n} = 1$. However, in the proof they make use of the equation $\lim_{n \rightarrow \infty} (1 - P_{n,k}) = \lim_{n \rightarrow \infty} \prod_{i=1}^{2n/k} (1 - p_i)$ which indirectly implies that they considered $1 - P_{n,k} = \prod_{i=1}^{2n/k} (1 - p_i)$ to be valid, where the probability p_i of reaching $y = k$ in the i^{th} section is at least p_1 . That, nevertheless, is incorrect (for a counter-example can easily be obtained). In Section 4 we provide an accurate proof of this proposition.

2. The sock matching theorem

Theorem 2.1. (*The Sock Matching Theorem - I alternative*)

The sequence $B_{n,k}$ whose n^{th} term represents the number of Dyck paths of order n which hit or cross the line $y = x + k$ is determined by the following recurrence formula:

$$B_{n,k} = \sum_{i=1}^n (B_{i-1,k-1} C_{n-i} + C_{i-1} B_{n-i,k} - B_{i-1,k-1} B_{n-i,k}). \quad (3)$$

PROOF. (The first one)

Let (i, i) be the first point on the line $y = x$ which the Dyck path visits after $(0, 0)$. Further, we take three possibilities into consideration: the line hits $y = x + k$ before (i, i) (Case 1), the line hits $y = x + k$ after (i, i) (Case 2), and the line hits $y = x + k$ both before and after (i, i) (Case 3).

The number of paths in the first case is $B_{i-1,k-1} C_{n-i}$. Namely, the number of ways to hit $y = x + k$ between $(0, 0)$ and (i, i) without hitting $y = x$ is the same as the number of ways to get from $(0, 1)$ to $(i-1, i)$ hitting $y = x + k$ but not crossing $y = x + 1$, which is $B_{i-1,k-1}$ and the number of ways to get from (i, i) to (n, n) without crossing $y = x$ is C_{n-i} .

Similarly, the numbers in the second and third case are $C_{i-1} B_{n-i,k}$ and $B_{i-1,k-1} B_{n-i,k}$, respectively. \square

PROOF. (The second one)

There is, however, yet another approach which may be taken in order to obtain the formula (3) which makes use of the recurrence relation satisfied by the numbers $A_{n,k}$ derived in [1]:

$$A_{n,k+1} = A_{n-1,k+1} A_{0,k} + A_{n-2,k+1} A_{1,k} + \dots + A_{0,k+1} A_{n-1,k}, \quad \text{for } n \geq 1, \quad k \geq 0; \quad (4)$$

with the initial conditions for $A_{n,0} = 0$, when $n \geq 1$ and for $A_{0,k} \stackrel{\text{def}}{=} 1$, when $k \geq 0$. To be more specific, as

$$B_{n,k} = C_n - A_{n,k-1}, \quad \text{for } n \geq 1 \text{ and } k \geq 1 \quad (5)$$

(evidently, $B_{n,1} = C_n$) making the necessary substitutions in (4) and with the use of the well-known recurrence relation for Catalan numbers ($C_0 = 1$, $C_{n+1} = \sum_{i=0}^n C_i C_{n-i}$, for $n \geq 0$) we obtain the desired relation. \square

Theorem 2.2. (*The Sock Matching Theorem - II alternative*)

$$B_{n,k} = \sum_{j=0}^{n-1} (B_{j,k} C_{n-j-1} + C_j B_{n-j-1,k-1} - B_{j,k} B_{n-j-1,k-1}). \quad (6)$$

PROOF. Similarly to the previous alternative we take the point (i, i) into consideration, only this time as the last point on the line $y = x$ that the Dyck path visits before (n, n) . Seen from this perspective, the corresponding numbers for the three cases would be exactly the values $B_{j,k} C_{n-j-1}$, $C_j B_{n-j-1,k-1}$ and $B_{j,k} B_{n-j-1,k-1}$.

By the way, the proof for the second formula (6) could, obviously, be obtain from (3) by a fairly simple substitution: $i = n - j$. \square

3. The explicit formula for $B_{n,k}$

We now give the explicit expression for the values of $B_{n,k}$.

Theorem 3.1.

$$B_{n,k} = \sum_{j=1}^{\lfloor \frac{n+1}{k+1} \rfloor} \binom{2n+2}{n+1-j(k+1)} - 4 \sum_{j=1}^{\lfloor \frac{n}{k+1} \rfloor} \binom{2n}{n-j(k+1)} \quad (7)$$

PROOF. (The first one) Recall that the value $A_{n,t}$ for the lower bound $h = 0$ is known from (1). Further, setting the upper bound to be $t = k - 1$ it follows directly from (5) that

$$B_{n,k} = \frac{1}{n+1} \binom{2n}{n} - \sum_{j \geq 0} \left[\binom{2n}{n-j(k+1)} - \binom{2n}{n-j(k+1)-1} - \binom{2n}{n-j(k+1)-k} + \binom{2n}{n-(j+1)(k+1)} \right] \quad (8)$$

After some minor algebraic simplifications of the above expression we have

$$B_{n,k} = \sum_{j \geq 1} \left[\binom{2n}{(n+1)-j(k+1)} - 2 \binom{2n}{n-j(k+1)} + \binom{2n}{(n-1)-j(k+1)} \right], \quad (9)$$

which coincides with the formula presented in [1] for the number of planted plane trees with $n+1$ nodes whose height is grater than k . Now, since

$$\binom{2n+2}{n+1-j(k+1)} = \binom{2n+1}{n-j(k+1)} + \binom{2n+1}{n+1-j(k+1)} = \binom{2n}{n-1-j(k+1)} + 2 \binom{2n}{n-j(k+1)} + \binom{2n}{n+1-j(k+1)},$$

the expression (7) is easily derivable from (9). \square

$B_{n,k}$	$k=1$	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$B_{1,k}$	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$B_{2,k}$	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$B_{3,k}$	5	4	1	0	0	0	0	0	0	0	0	0	0	0	0
$B_{4,k}$	14	13	6	1	0	0	0	0	0	0	0	0	0	0	0
$B_{5,k}$	42	41	26	8	1	0	0	0	0	0	0	0	0	0	0
$B_{6,k}$	132	131	100	43	10	1	0	0	0	0	0	0	0	0	0
$B_{7,k}$	429	428	365	196	64	12	1	0	0	0	0	0	0	0	0
$B_{8,k}$	1430	1429	1302	820	336	89	14	1	0	0	0	0	0	0	0
$B_{9,k}$	4862	4861	4606	3265	1581	528	118	16	1	0	0	0	0	0	0
$B_{10,k}$	16796	16795	16284	12615	6954	2755	780	151	18	1	0	0	0	0	0
$B_{11,k}$	58786	58785	57762	47840	29261	13244	4466	1100	188	20	1	0	0	0	0
$B_{12,k}$	208012	208011	205964	179355	119438	60214	23276	6854	1496	229	22	1	0	0	0
$B_{13,k}$	742900	742899	738804	667875	477179	263121	113620	38480	10075	1976	274	24	1	0	0
$B_{14,k}$	2674440	2674439	2666248	2478022	1877278	1116791	528840	200655	60606	14301	2548	323	26	1	0
$B_{15,k}$	9694845	9694844	9678461	9180616	7303360	4637476	2375101	990756	336168	91756	19720	3220	376	28	1

Tabular 1 Numerical values of $B(n, k)$ for small values of n and k

PROOF. (The second one)

Let us now commence from the formula (2). By applying $\sin^2 \alpha = 1 - \cos^2 \alpha$ on it we obtain the following

$$A_{n,t} = \frac{4^{n+1}}{t+2} \left(\sum_{1 \leq j \leq \frac{t+1}{2}} \cos^{2n} \left(\frac{j\pi}{t+2} \right) - \sum_{1 \leq j \leq \frac{t+1}{2}} \cos^{2n+2} \left(\frac{j\pi}{t+2} \right) \right), \quad n \geq 1. \quad (10)$$

Further, using one of the notations for the representation of trigonometric power sums, namely the one with binomial coefficients, from [3], we have

$$\sum_{j=0}^{N-1} \cos^{2m} \left(\frac{j\pi}{N} \right) = 2^{1-2m} N \left(\binom{2m-1}{m-1} + \sum_{p=1}^{\lfloor \frac{m}{N} \rfloor} \binom{2m}{m-pN} \right), \text{ for } m \geq N, \quad m, N \in \mathbb{N}. \quad (11)$$

Now, making the necessary substitutions, i.e. $N = t + 2$ and for m at first $m = n$ and then $m = n + 1$, a brief simplification process leads to

$$A_{n,t} = 4 \left[\binom{2n-1}{n-1} + \sum_{j \geq 1}^{\lfloor \frac{n}{t+2} \rfloor} \binom{2n}{n-j(t+2)} \right] - \left[\binom{2n+1}{n} + \sum_{j \geq 1}^{\lfloor \frac{n+1}{t+2} \rfloor} \binom{2n+2}{n+1-j(t+2)} \right]. \quad (12)$$

Once again, utilising (5) and yet again simplifying the obtained expression we eventually come to the desired formula (7) \square

4. Asymptotic behavior

Theorem 4.1. *The probability $P_{n,k}$ of reaching a given fixed k approaches 1 as n approaches infinity, i.e.*

$$\lim_{n \rightarrow \infty} P_{n,k} = \lim_{n \rightarrow \infty} \frac{B_{n,k}}{C_n} = 1.$$

PROOF. Exploiting (5) and (2) some more we have

$$\lim_{n \rightarrow \infty} P_{n,k} = \lim_{n \rightarrow \infty} \frac{B_{n,k}}{C_n} = 1 - \lim_{n \rightarrow \infty} \frac{A_{n,k-1}}{C_n} = 1 - \lim_{n \rightarrow \infty} \frac{\frac{1}{k+1} \sum_{1 \leq j \leq \frac{n}{k+1}} 4^{n+1} \sin^2 \left(\frac{j\pi}{k+1} \right) \cos^{2n} \left(\frac{j\pi}{k+1} \right)}{\frac{1}{n+1} \frac{(2n)!}{n!n!}}$$

The first summand in the above sum is the dominant one thus

$$A_{n,k-1} \sim \frac{1}{k+1} 4^{n+1} \sin^2 \left(\frac{\pi}{k+1} \right) \cos^{2n} \left(\frac{\pi}{k+1} \right), \text{ for fixed } k, n \rightarrow \infty.$$

Consequently,

$$\lim_{n \rightarrow \infty} P_{n,k} = 1 - \lim_{n \rightarrow \infty} \frac{\frac{1}{k+1} 4^{n+1} \sin^2 \left(\frac{\pi}{k+1} \right) \cos^{2n} \left(\frac{\pi}{k+1} \right)}{\frac{1}{n+1} \frac{(2n)!}{n!n!}}.$$

Applying the Stirling's approximation we have

$$\lim_{n \rightarrow \infty} P_{n,k} = 1 - \lim_{n \rightarrow \infty} \left[\frac{4(n+1)}{k+1} \sqrt{\pi n} \sin^2 \left(\frac{\pi}{k+1} \right) \cos^{2n} \left(\frac{\pi}{k+1} \right) \right].$$

Bearing in mind that $\cos^{2n} \left(\frac{\pi}{k+1} \right)$ approaches zero faster than $1/((n+1)\sqrt{n})$, it follow immediately that $\lim_{n \rightarrow \infty} P_{n,k} = 1$. \square

Acknowledgments

We wish to express our sincerest gratitude towards Dragan Stevanović for pointing out references [6] and [3].

References

- [1] N. G. de Bruijn, D. E. Knuth, S. O. Rice, The average height of planted plane trees, in *Graph theory and computing*, edited by R.C.Read, Academic Press, New York, 1972. MR 58 # 21737 Zbl 0247.05106, 15-22.
- [2] S. Gilliland, C. Johnson, S. Rush, D. Wood, The sock matching problem, *Involve*, 7 (5) (2014), 691–697.
- [3] C. M. da Fonseca, M. L. Glasser, V. Kowalenko, Basic trigonometric power sums with applications, arXiv:1601.07839v1 [math.NT] 28 jan 2016.
- [4] A. Ilić, A. Ilić, On the number of restricted Dyck paths, *Filomat*, 25:3 (2011), 119–201.
- [5] W. Panny, H. Prodinger, The expected height of paths for several notions of height, *Studia Scientiarum Mathematicarum Hungarica* 20 (1985), 119–132.
- [6] H. Prodinger, The number of restricted lattice paths revisited, *Filomat*, 26 (6) (2012), 1133–1134.
- [7] R. P. Stanley, *Enumerative Combinatorics*, Vol. I, Cambridge University Press, Cambridge, 2002.
- [8] R. P. Stanley, *Catalan Addendum to Enumerative Combinatorics*, Volume 2, version of 25 May 2013, <http://www-math.mit.edu/~rstan/ec/catadd.pdf>